Comment

Some Observations on ACE Methodology

ANDREAS BUJA and ROBERT E. KASS*

1. INTRODUCTION

Breiman and Friedman's development of the ACE algorithm and associated methodology will surely be seen as a milestone in the theory of nonparametric regression. The authors have not only provided theoretical analysis of ACE applied to random variables, they have also incorporated fast smoothing techniques to make the method work efficiently on finitely many observations, and they have tested their implementation on both simulated and real data. The statistical community has been presented with a tool that is conceptually simple, mathematically elegant, and computationally economical, as well as being useful in its demonstrated ability to uncover nonlinear relationships that might otherwise be missed.

As a data-analytic technique, ACE could be viewed as a generalization of alternating least squares (ALS) developed by Young (1981), DeLeeuw (1983), and their co-workers. Applications to several important classes of problems have appeared, primarily in the psychometric literature. This methodology is itself an outgrowth of "optimal scoring" of categorical data, which goes back at least to Fisher (1941), who posed and solved the problem: "Given a two-way table of non-numerical observations we may ask what values, or scores, shall be assigned to them in order that the observations shall be as additive as possible" (p. 283).

It is important that, in addition to the polynomial fitting methods used for quantitative data by the psychometricians, a variety of expensive smoothers, such as splines, could be used together with methods of numerical linear algebra. The Breiman–Friedman implementation, however, offers an advance by incorporating the fast smoother technology of Friedman and Stuetzle (1982), thereby lowering cost and increasing applicability.

The goal of ACE is similar to that of Fisher's optimal scoring method; it finds the transformations that make the relationship of $\theta(Y)$ to the $\varphi_j(X_j)$'s as linear as possible, where departure from linearity is measured by expected squared error relative to total variance. If ACE is applied to the transformed response $\tilde{Y} = \theta(Y)$ and the normalized carriers $\tilde{X}_j = \varphi_j(X_j)/\|\varphi_j(X_j)\|$, then the optimal transformations will be $\tilde{\theta}(Y) = \tilde{Y}$ and $\tilde{\varphi}_j = \beta_j \tilde{X}_j$, where $\beta_j = \|\varphi_j(X_j)\|$. That is, linear regression will be optimal and $E(\tilde{Y} \mid \tilde{X}_1, \ldots, \tilde{X}_p) = \sum \beta_j \tilde{X}_j$. Characterization problems concerning "linearity of regression" have been around a long time, a reference being Kagan et al. (1973, pp. 10–12). An important feature of ACE is that optimal transformations are not the only ones that produce linear regressions. In general, there exist many sets of stationary values of the criterion, which correspond to eigenfunctions associated with different eigenvalues of ACE's iterated conditional expectation operator in Hilbert space, and each of these shares the property of linearity of regression. Indeed, as we discuss in Section 2, the suboptimal eigenfunctions can sometimes be of greater interest than the optimal ones.

Another way to view ACE is as a nonparametric alternative to power transformations based on the assumption of Normality, having been developed in the spirit of exploration, rather than inference. A basic element of ACE is its reversal of the standard practice of treating stochastic predictors as if they were fixed and had independent additive errors attached: ACE eliminates the additive error structure and replaces it with a joint distribution of response and predictor variables. We will comment further on the distinction between stochastic and investigator-determined predictors in Section 3. For now, and in Section 2, we will take as unobjectionable the assumption, which is essential to the theory of ACE, that $(X_1, \ldots, X_p, Y)$ has a multivariate distribution (and the marginal distributions are nondegenerate). In this context it is worth emphasizing the result cited near the end of Breiman and Friedman's Section 1: If marginal transformations to joint Normality exist, then ACE will find them. We notice in addition that if *monotonic* transformations to joint Normality exist, transformation of each marginal distribution to Normality will also produce the joint Normal distribution. This strong assumption is the basis for the common practice of examining and then transforming each variable separately; when it holds, ACE, too, will succeed (albeit somewhat less efficiently), and the transformations will be roughly the same. The results from ACE deviate from those of available parametric methodology when the distribution after transformation is non-Normal or the optimal transformations are nonmonotonic. Here ACE gains its great nonparametric advantage but, as we next discuss, it is important to keep in mind that what is "optimal" need not be desirable.

2. OPTIMAL AND SUBOPTIMAL EIGENFUNCTIONS

If we consider the simplest class of non-Normal joint distributions, those that are elliptically symmetric, an interesting anomaly appears. When the association among the variables is sufficiently weak, optimal transformations may break the symmetry and thus might be judged misleading. Taking, for simplicity, the bivariate distribution of a pair of random variables (X, Y), in the extreme case of spherical symmetry X and Y are independent if and only if they are jointly Normal. From property 2 of Breiman and Friedman's Section 1 it follows that the optimal correlation is zero if and only if X and Y are Normal. Thus the optimal transformations of X and Y are different from the identity whenever the spherical bivariate distribution is non-Normal. As an example, consider the uniform distribution on

* Andreas Buja is Assistant Professor, Department of Statistics, University of Washington, Seattle, WA 98195. Robert E. Kass is Assistant Professor, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213. Research was supported by National Science Foundation Grants MCS-8304234 and MCS-8301831.

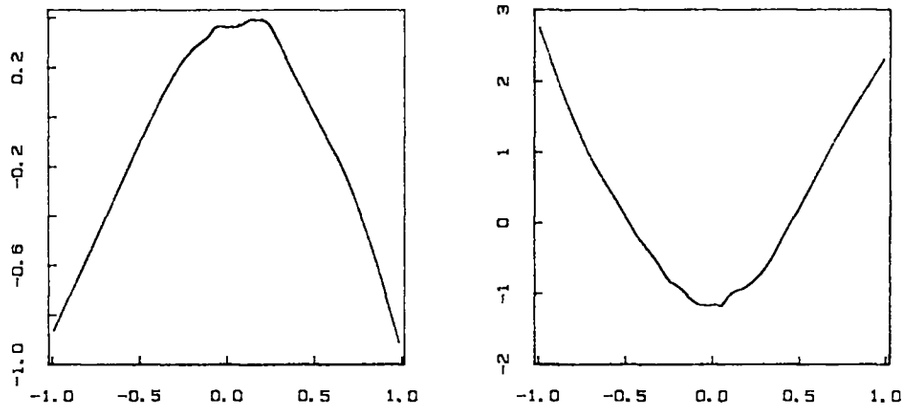

Figure 1. ACE Transformations of X (left) and Y (right) for Correlation .0.

the unit disk $x^2 + y^2 < 1$. It turns out that the optimal transformations φ and θ are parabolas symmetric about the vertical axis, and simulations show that ACE finds these in finite data (see our Figure 1). The optimal correlation is $\frac{1}{3}$ for the populations.

Proceeding a step further, one can apply linear transformations to (X, Y) to produce a class of distributions that are uniform on elliptical disks and have correlations ranging from 0 to 1. The optimal transformations for the distributions with correlation close to 1 are the identity, whereas those with correlation close to 0 are again parabolas; the transition from one case to the other is abrupt and occurs for correlation $\frac{1}{4}$. (A derivation of this theoretical fact will appear somewhere else.)

These remarks apply to the underlying populations rather than finite samples. As a demonstration for the finite sample case, we included some plots of ACE transforms that were generated with Breiman and Friedman's program (see Figures 1–5). The data sets were obtained by linear transformation of 391 pseudorandom points from a uniform distribution on the unit disk in $\mathbf{R}^2$. The linear transformations were chosen so as to yield pseudorandom samples from uniform distributions on elliptic disks with correlation .0 (no transformation), .22, .28, .34, and .5. One notices that the transforms change shape fairly quickly in the range from .22 to .34. The particular pseudorandom sample at hand resulted in empirical correlations somewhat below the theoretical ones (.19, .25, .32 instead of .22, .28, .34), which may partly account for the qualitative jump from Figure 3 to Figure 4 rather than from Figure 2 to Figure 3 as expected from theory. Nevertheless, it seems that the finite-sample ACE algorithm reflects the behavior of ACE at the underlying population. We would like to mention that in our experience for correlation close to $\frac{1}{4}$, the finite-sample ACE algorithm occasionally produces transforms of cubic shape.

How should we understand this behavior? From a mathematical point of view, ACE finds eigenfunctions of largest eigenvalues by applying the power method to the iterated conditional expectation operators $E(E(\cdot \mid Y) \mid X)$ for $\varphi(X)$ and $E(E(\cdot \mid X) \mid Y)$ for $\theta(Y)$, the largest eigenvalue being the squared optimal correlation (compare Theorem 5.3). The uniform distributions on elliptical disks all have the same systems of eigenfunctions, which are polynomials, but the eigenvalues vary with the degree of ellipticity. It turns out that the polynomial of order 1—that is, the identity—has the largest eigenvalue for strong ellipticity (corresponding to correlation greater than $\frac{1}{4}$), but it is overtaken by the second-order polynomial for weak ellipticity (correlation less than $\frac{1}{4}$). As the ellipticity weak-

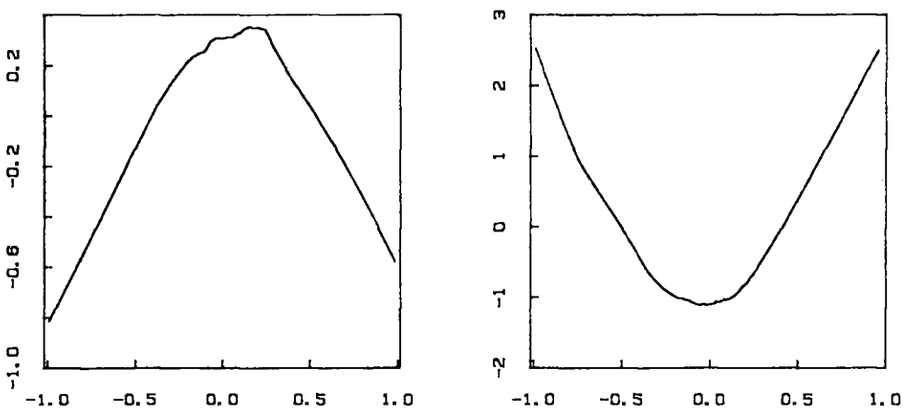

Figure 2. ACE Transformations of X (left) and Y (right) for Correlation .22.

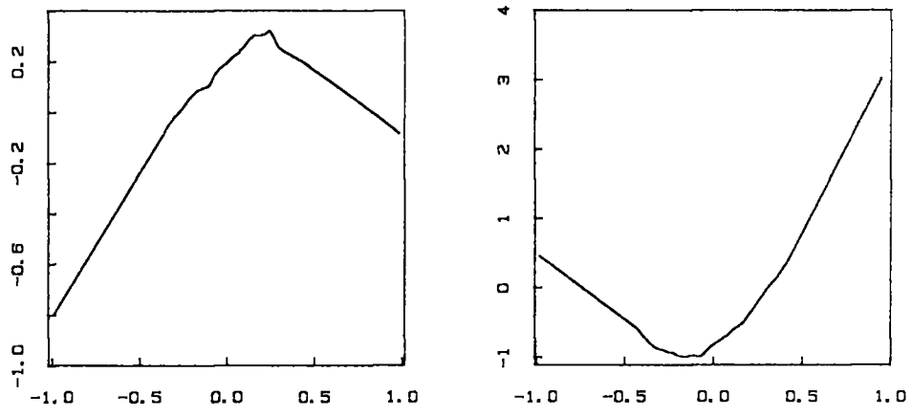

Figure 3. ACE Transformations of X (left) and Y (right) for Correlation .28.

ens, axial symmetry is approached and this drives the eigenvalues of odd-order polynomials to zero. In Figure 6, the square root of the eigenvalues as a function of correlation (ellipticity) is plotted for the eigenpolynomials of degrees 1–3. Notice that not only does the second-degree polynomial have the largest eigenvalue for correlation below $\frac{1}{4}$, but the eigenvalue of the third-degree polynomial is suspiciously close to the eigenvalue of both the first- and second-degree polynomial in a neighborhood of $\frac{1}{4}$. This is probably why we occasionally observed cubic behavior in the ACE transforms of finite samples.

A background for this type of phenomenon is provided by perturbation theory (see Kato 1984, pp. 63–74), which deals with the dependence of linear operators and their spectral decompositions on a real or complex parameter. Here and in numerical analysis of eigenproblems (Parlett 1980, pp. 14–15), one knows that eigenvalues depend continuously on the operator, but eigenfunctions do not. In our context this means that the optimal correlation of two close distributions will be close, but the ACE transforms may look qualitatively different. A case in point is the preceding example, where a continuous change in ellipticity leads to a jump at the critical correlation $\frac{1}{4}$, but more serious discontinuities are possible (Kato 1984, pp. 64, 72). It is always multiplicity (also called degeneracy) of eigenvalues that gives rise to this phenomenon. In quantum physics, degeneracy is often the generic case, and perturbations (e.g., of an atom by an external magnetic field) are known to split up spectra (the Zeeman effect).

The occurrence of nonmonotonic optimal transformations in the preceding example should not be interpreted as revealing remarkable relationships in the data. Given the situation in the spherical case, an obvious question is whether Normal distributions are the only elliptically symmetric bivariate distributions for which the optimal transformations are the identity for all degrees of ellipticity as measured by the correlation of the untransformed data. If the preceding uniform distributions are examples for short-tailed cases, one might consider at the other end some simple heavy-tailed examples, such as spherical and elliptical t distributions. We are in the process of examining such examples.

The possibility of finding many relevant eigenfunctions is not limited to the lower range of optimal correlations, as is seen if we choose as a distribution for (X, Y) the degenerate uniform on the diagonal of the unit square in the plane. Here we have $P(X = Y) = 1$, an exact relationship. This example is, strictly speaking, outside of the Breiman–Friedman setup, due to degeneracy as a measure (see Assumption 5.2), yet it

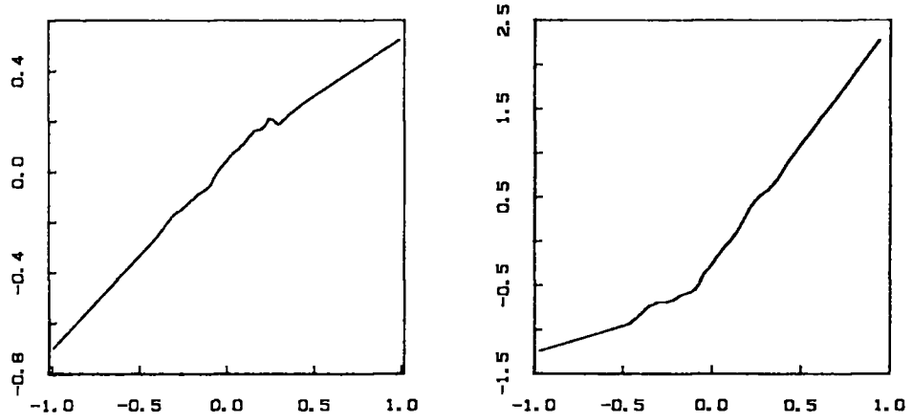

Figure 4. ACE Transformations of X (left) and Y (right) for Correlation .34.

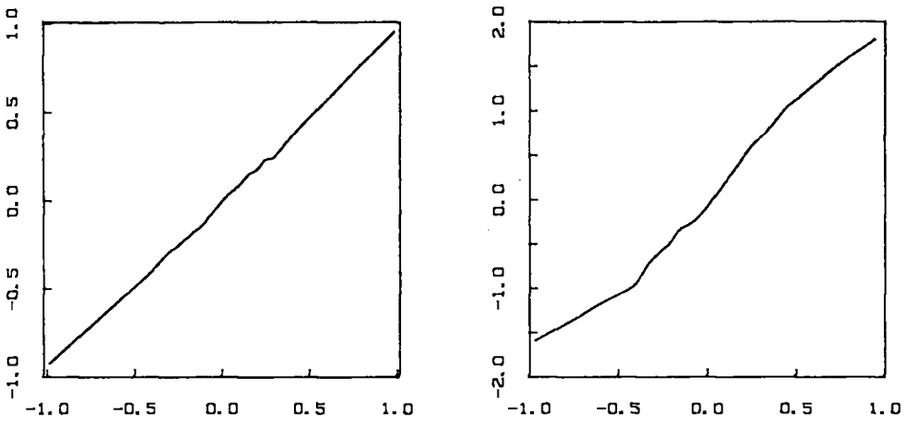

Figure 5. ACE Transformations of X (left) and Y (right) for Correlation .5.

indicates a seemingly strange behavior of ACE: There do exist uncountably many transformations to optimal correlation 1. Any pair (θ, φ) of transforms will do if $\theta = \varphi$ and if θ, φ are nonconstant with probability 1 on the unit interval. This example is extreme and has little relevance for data analysis. The following, however, is a real possibility: The joint distribution of $(Y, X_1, \ldots, X_p)$ might approximately satisfy more than one additive equation of the form $\theta(Y) = \sum \varphi_j(X_j)$ and thus cluster around a manifold of codimension more than 1 in $\mathbf{R}^{p+1}$. Such a situation leads to either multiple optimal transformations or a spectrum that features several eigenvalues bunching together at the upper end, and they may even be arbitrarily close to 1. The point here is that we might not only encounter multiple sets of predictor transforms $\varphi_j(X_j)$, but response transforms $\theta(Y)$ as well. In this respect, ACE differs from the untransformed linear models, where multiple fits are always due to dependencies among the predictors alone.

In these situations a possible safeguard would be to obtain several of the largest eigenvalues and their eigenfunctions. Examination of the upper end of the spectrum would indicate the stability of the optimal transformations. A single largest eigenvalue fairly close to 1 would give some reassurance that the additive model produced by ACE is appropriate. However, if there did exist more than one eigenvalue close to 1, further scrutiny would be required, and the possibility of describing the data by more than one additive equation would have to be considered.

A related concern arises in the following situation: Let the distribution of (X, Y) fall apart into two natural clusters in diagonally opposite quadrants; that is, there should exist thresholds a and b such that $P(X \leq a, Y \leq b)$ and $P(X > a, Y > b)$ are both nonzero and sum to 1. For such a distribution, the optimal correlation happens to be 1! Optimal transformations that achieve it can be obtained by letting $\theta(Y)$ map the sets $\{Y \leq b\}$ and $\{Y > b\}$ onto different constants, and similarly for $\varphi(X)$ with the sets $\{X \leq a\}$ and $\{X > a\}$. (We owe this observation to Charles Stone.) Hence ACE discovers that the data can be lumped together by cuts along the X and Y axes, and it uses this in the search for optimal marginal transformations. The finite-sample version of ACE based on the Friedman–Stuetzle smoothers will produce a variant of this result due to averaging over windows that extend a fraction (e.g., 15%) to the left as well as to the right of a given point. Thus the jumps at a and b in φ and θ, respectively, will be smeared out to steep but continuous stretches around the threshold values. Owen (1983, pp. 19–23) informed us that this phenomenon actually occurred in real data when he applied ACE in a time-series context, and the ACE transformations he found behaved as we described here. It seems to us that such simultaneous lumping in X and Y calls for a special treatment, but we do not yet have any suggestions. A perplexing fact is that this lumping effect occurs regardless of the distribution of (X, Y) in the two quadrants. For example, it may be nicely clustering around two fragments of manifolds so that ACE would have to pick up the manifold structure in its suboptimal solutions, which leads us once more to the recommendation of looking at suboptimal eigenfunctions.

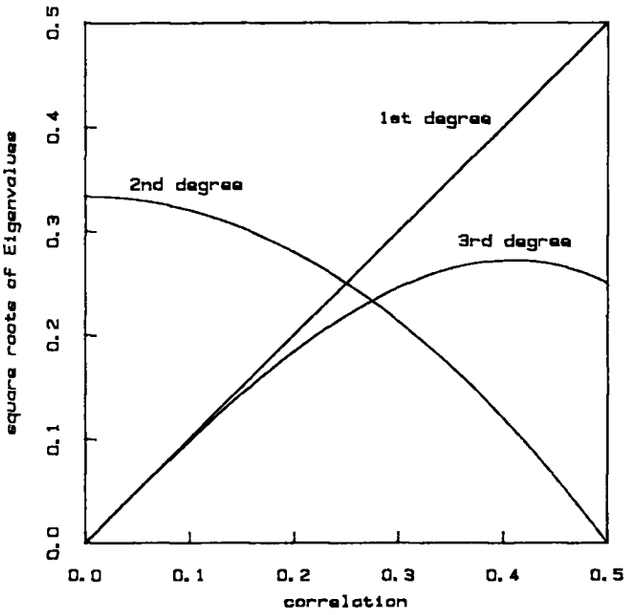

Figure 6. Square Roots of the Eigenvalues of First-, Second-, and Third-Order Eigenpolynomials as Functions of the Correlation of X and Y.

3. ACE AND PREDICTOR MODELS

Like parametric regression methods based on the assumption of joint Normality of predictors and responses, ACE blurs the distinction between stochastic and nonstochastic predictor variables. The conventional methods are inferential and rely on Fisherian or Bayesian principles to justify conditioning on stochastic predictors. ACE, on the other hand, does away with the additive error structure of predictor models, using instead a nondegenerate joint distribution. Thus ACE produces transformations that satisfy not only the additive equation

$$E(\theta(Y) \mid X_1, \ldots, X_p) = \sum \varphi_j(X_j),$$

but also the dual relation

$$E(\sum \varphi_j(X_j) \mid Y) = \rho\theta(Y),$$

where ρ is the optimal multiple correlation. This pair of equations forms an eigenproblem characterizing stationarity under ACE iteration, and it displays the symmetric treatment of response and carriers by ACE.

To some, this feature of ACE may seem unnatural, and the practical consequence, mentioned by Breiman and Friedman toward the end of Section 1, is that when data are generated from a predictor model

$$\theta(Y) = \sum \varphi_j(X_j) + \varepsilon,$$

ACE does not necessarily find the transformations $\varphi_1, \ldots, \varphi_p, \theta$. One situation in which ACE would clearly produce undesirable results was described earlier: If the design were such that the predictor values clustered into clearly separated subpopulations and the error ε were small relative to the separation, then ACE would produce lumps and the interesting transformations might correspond to suboptimal eigenfunctions. From its possible failure to produce desirable predictor-model transformations, we would not conclude that ACE is somehow flawed, but it does appear that there is more to be learned about the sensitivity of ACE to the distribution of the carriers. This is an issue especially for designed experiments and data from samples that poorly represent the population distribution of the carriers. On the other hand, in the difficult but common situation in which scientific investigation requires analysis of well-sampled observational data with little guidance from a predictive theory, the inferential approach based on predictive models is problematic and the powerful exploratory ACE technology should be especially useful.

4. INFERENCE, DIAGNOSTICS, AND ROBUSTNESS

One aspect of the problem of assigning scores discussed by Fisher was that of assessing variability. He noted that the complexity of the problem limited the applicability of the notion of standard error, and he proceeded to develop as an alternative a significance test to assess departure from a particular choice of scores. The generalization of a choice of scores is, of course, a choice of transformations. The transformations produced by ACE will probably be most often used as rough suggestions: Practitioners will interpret the plots as approximations to elementary functions such as logarithms, exponentials, powers, or trigonometric functions, and they will then build linear models based on these. Inference from the resulting models, including statements concerning possible inclusion of particular carriers in the model, would be difficult and treacherous. The most useful variabilities to assess would be those of the transformations themselves, although in this general context, as in optimal scoring, inversion to significance tests may again be much more tractable. It would also be nice to be able to judge the separation of several large eigenvalues, discussed in Section 1 here, against background variation. Perhaps, as Breiman and Friedman suggest (in Section 4), the bootstrap will be helpful in solving these problems.

There is also the possibility of developing diagnostics for use in conjunction with ACE. Although there may be goals that are analogous to those of diagnostics for linear regression, there will be important differences. In the case of colinearity, for example, in linear regression the existence of an exact colinearity of the form $\Sigma\, a_j \cdot X_j = 0$ implies that with any set of least-squares coefficients β_j, the set of numbers $\beta_j + a_j$ provides a least-squares solution. On the other hand, nonlinear relations among carriers do not necessarily produce infinitely many solutions. When applying ACE, however, the existence of a relation of the form $\Sigma\, \psi_j(X_j) = 0$ implies that with any set of optimal predictor transformations $\varphi_j(X_j)$, the transformations $\varphi_j(X_j) + \psi_j(X_j)$ are optimal as well. Thus it might be useful to develop methods for estimation of additive implicit equations of the form $\Sigma\, \psi_j(X_j) \approx 0$, in analogy to estimation of smallest principal components of the carriers in linear regression. One of us (Buja) has begun working on this together with Werner Stuetzle.

Finally, let us briefly mention the issue of robustness. Notice that although the theory presented by Breiman and Friedman is a second-order theory, based on the notions of correlation and conditional expectation as projections in L_2 space, the implied existence of second moments refers not to the variables themselves but to the *transformations* of them. Hence it makes sense to consider even spherical and elliptical Cauchy distributions (the only problem being that the ACE algorithm must not be initialized with the identity transforms, since they are not elements of L_2 space). Nevertheless, those who think most easily in terms of models may wonder what there is to recommend the use of correlation with non-Normal data. In more practical terms, there may be problems of robustness in using ACE, and with worries about effects of outliers in mind, one might wonder about the advisability of using a criterion based on squared error. It might be possible to find another useful sense in which an average departure from linearity could be minimized, but it is far easier to make this glib remark than to formulate the problem in such a way that progress could be made while retaining such advantages of the current ACE algorithm as the low computational cost. A simple ad hoc approach would be to use a cheap robust smoother, but it is quite possible that the overall statistical performance of ACE would be adversely affected.

5. CONCLUSIONS

There is ample evidence that ACE, in its present form, is already a greatly useful tool for data analysis. We have de-

scribed situations in which interpretation of ACE results requires caution, and we have mentioned some possibilities for handling these situations and extending ACE technology.

ADDITIONAL REFERENCES

DeLeeuw, J. (1983), "The Gifi System of Non-Linear Multivariate Analysis," in *Data Analysis and Informatics*, eds. E. Diday and L. Lebart, Amsterdam: North-Holland.

Fisher, R. A. (1941), *Statistical Methods for Research Workers* (8th ed.), New York: G. E. Stechert (14th ed., New York: Hafner Press).

Kagan, A. M., Linnik, Yu. V., and Rao, C. R. (1973), *Characterization Problems in Mathematical Statistics*, New York: John Wiley.

Kato, T. (1984), *Perturbation Theory for Linear Operators* (2nd ed.), New York: Springer-Verlag.

Owen, A. (1983), "Optimal Transformations for Autoregressive Time Series Models," Technical Report 020, Project Orion, Stanford University, Dept. of Statistics.

Parlett, B. N. (1980), *The Symmetrical Eigenvalue Problem*, Englewood Cliffs, NJ: Prentice-Hall.

Young, F. W. (1981), "Quantitative Analysis of Qualitative Data," *Psychometrika*, 46, 357–388.

Comment

The ACE Method of Optimal Transformations

E. B. FOWLKES and J. R. KETTENRING*

1. INTRODUCTION

The idea that one can achieve "optimal" transformations of the variables in a regression problem by repeated application of a two-variable smoothing algorithm is fascinating. In developing this idea, Breiman and Friedman have brought to bear, in elegant fashion, the statistical theory of maximal correlation, the mathematics of Hilbert spaces, and the ACE algorithm for translating the theory into practice.

ACE is powerful medicine. It has the ability to uncover very general transformations, and it seems to find them when they are needed. But it also appears to have some unwanted side effects in cases where there is no really interesting structure to be found. People who will use ACE in practice—and that includes us—will need a careful characterization of what the algorithm does under different scenarios.

We will focus our detailed comments on the use of ACE in practice, based on our own limited experiences; some empirical properties of ACE, based on a small set of experiments; and possible adaptations of ACE to other problems in multivariate analysis.

2. ACE AS A DATA-ANALYTIC TOOL FOR REGRESSION

ACE promises to be an important new tool for the data analyst in carrying out regression analyses. However, it should be studied and used in the context of the ever-growing collection of modern tools for regression such as robust procedures, diagnostic techniques for identifying influential points, and so forth. As data analysts, we use modern regression tools in a flexible manner, moving from one to another and comparing the salient features of the data that may be revealed.

We illustrate how we have made use of ACE in a real example. It concerns relating, for a particular telephone switching entity, a measure of the total call load for an interval of time (measured in CCS, or 100 call seconds, where 1 CCS is any combination of calls that accounts for 100 seconds) to five different types of service (residence local, residence metro, business metro, business local, and coin, measured by the number of associated telephone lines). A large number of data-analytic techniques were used on these data, but we shall only consider those that have a relationship with ACE. Figure 1 shows scatterplots of the raw data for CCS versus residence local, CCS versus residence metro, and residence local versus residence metro. The three scatterplots are typical of all possible scatterplots of the response, CCS, and the explanatory variables. Striking features of the scatterplots include the large density of points near (0, 0), the change in variability for increasing values of the variables, and the one high leverage point (not an error). Experience suggested that a reasonable first step in analyzing these data would be to transform the response, explanatory variables, or both; and we decided to take logarithms of all variables. Figure 2 shows scatterplots of the same variables considered in Figure 1 after the log transformations had been made. The problems of great differences in density and variability have been alleviated, and the effect of the leverage point has been diminished. There is a hint of curvilinearity in the log CCS versus log residence metro plot. We applied ACE to the transformed data in the spirit of a check on what had been suggested by examining the scatterplots. If the log transformations were reasonable, the transformations indicated by ACE should be linear. Figure 3 shows the results. The ordinate scales have been equalized in order to facilitate a judgment of the relative contributions of the different explanatory variables. The figure indicates that the transformations for CCS and residence local are indeed linear, but the one for residence metro is systematically curved. The figure also indicates that residence local, with its wider range of transformed values, is a more

* E. B. Fowlkes and J. R. Kettenring are Members of Technical Staff, Statistics Research Group, Bell Communications Research, Morristown, NJ 07960. They thank R. Gnanadesikan and P. A. Tukey for their comments. The examples reported here were carried out using a version of the basic ACE program (dated October 6, 1983) and a version of the "super" smoother employed by this program (dated March 10, 1984).

Journal of the American Statistical Association
September 1985, Vol. 80, No. 391, Theory and Methods